Computer Architechture

Floating Point Discussion

Zhongyue Lin

linzhy@shanghaitech.edu.cn

1.IEEE754(32-bit, single-precision)

1	8	23
Sign	Exponent	Significand

Sign Bit: 0 positive, 1 negative

Exponent: Biased notation. Bias of single-precision is 127 ($2^{8-1} - 1$)

Read the exponent and subtract with bias. The reason is that we want easy comparisons of the exponent. (we don’t want extract exponent and decode and compare it in two’s complement.)

The value of bias is 0b01111111 (the first bit is a zero followed by all ones).

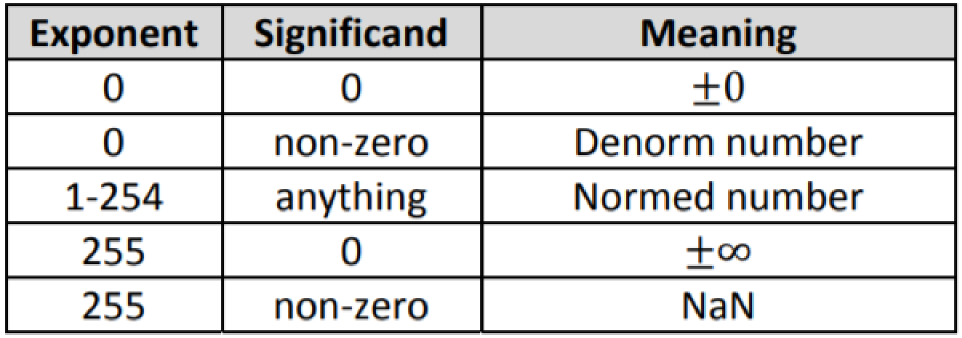
1.IEEE754(32-bit, single-precision)

1	8	23
Sign	Exponent	Significand

Significand: Implicit leading 1: 1.abcd… abcd are the bits in significands from left to right.

We want this implicit 1 for more representation range.

1.IEEE754(32-bit, single-precision)



Exponent	Significand	Meaning
0	0	±0
0	non-zero	Denorm number
1-254	anything	Normed number
255	0	±∞
255	non-zero	NaN

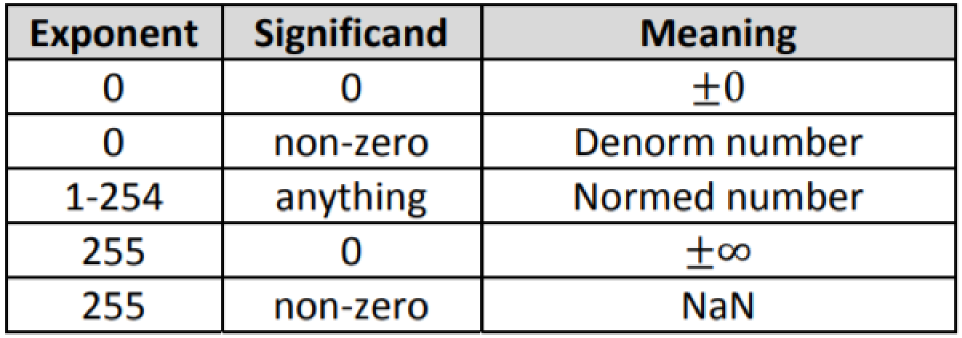
Normal Numbers: $(-1)^{Sign} * 2^{Exp-Bias} * 1.Significand_2$

Denorm: $(-1)^{Sign} * 2^{Exp-Bias+1} * 0.Significand_2$

1. Due to our implicit leading 1 in significands, there are holes in representing very small numbers, we need Denorm number with exponent all zeros.

2. There is no implicit leading 1 in denorm and there is a plus one in exponent.

1.IEEE754(32-bit, single-precision)



Exponent	Significand	Meaning
0	0	± 0
0	non-zero	Denorm number
1-254	anything	Normed number
255	0	$\pm\infty$
255	non-zero	NaN

Normal Numbers: $(-1)^{Sign} * 2^{Exp-Bias} * 1.Significand_2$

Denorm: $(-1)^{Sign} * 2^{Exp-Bias+1} * 0.Significand_2$

The smallest positive normed number you get is

$$2^{1-127} * (0b1.00 \ldots 00) = 2^{-126}$$

The smallest positive denorm number you get is

$$2^{0-127+1} * (0b0.00 \ldots 01) = 2^{-149}.$$

2.Git with Software

Fork (Mac && Win)

GitKraken (Mac && Win && Linux)

Questions?

上海科技大学

立志成才 报国裕民

2013

SHANGHAITECH UNIVERSITY